

## Completing the *English Vocabulary Profile*: C1 and C2 vocabulary

Annette Capel Freelance consultant based in Cambridge

### Abstract

*The English Vocabulary Profile is an online vocabulary resource for teachers, teacher trainers, exam setters, materials writers and syllabus designers. It offers extensive information about the Common European Framework of Reference (CEFR) levels of words, phrases, phrasal verbs and idioms, and currently includes just under 7,000 headwords. This article reports on the trialling and validation phase of the A1–B2 levels of the resource, as well as outlining the research and completion of the C1 and C2 levels. The project has followed a ‘can-do’ rationale, focusing on what learners actually know rather than prescribing what they should know, and is underpinned by up-to-date corpus evidence, including the 50-million word Cambridge Learner Corpus and the 1.2-billion word Cambridge English Corpus of first language use. At C1 and C2 levels, the English Vocabulary Profile describes both General and Academic English, and the additional sources used to research this area of language learning are described in the article. Polysemous words are treated in depth and the project has sought to determine which meanings of these important words appear to be acquired first; new, less frequent meanings often continue to be learned across all six CEFR levels. Phrases form another substantial part of the resource and this aspect has been guided by expert research (see Martínez 2011).*

**Keywords:** CEFR, corpus, idioms, phrases, polysemous words, word families

This article, written at the end of the compiling stage of the C1 and C2 levels of the English Vocabulary Profile (formerly known as the English Profile Wordlists), describes the work that has been carried out on the project since September 2010. (For information about the initial phase of the project, see Capel 2010.)

### 1. The English Vocabulary Profile

The decision to move away from the original working title of ‘Wordlists’ reflects how this project has grown, not just in terms of its coverage of vocabulary, to include many more phrases, phrasal verbs and idioms rather than just words, but also the way in which the current online resource has been developed to provide a fully interactive database, instead of being a static listing of vocabulary. Even now, when the compiling process has come to an end with the inclusion of C1 and C2 level data, the resource is not (and never will be) set in

stone. It represents the extent of the description that is currently achievable given the learner data and other sources that the project team has access to.

It goes without saying that the resource will need to be regularly monitored and refined, partly to keep it up to date but equally to ensure that it accurately reflects typical learner competence. As additional learner evidence becomes available in the form of spoken and non-exam written data – the Cambridge English Profile Corpus – and as more people use the resource and give their feedback on it, this community project will be honed and augmented.

Public access to the resource is currently limited to the A and B levels of the British-English and American-English versions, which have both been validated over a twelve-month period (see Section 3 below). At the time of writing (February 2012), the resource is available for free on the main English Profile website. It is hoped that the complete six-level resource will be available on general release in May 2012.

## 2. Coverage within the six levels of the English Vocabulary Profile

For CEFR levels A1 to B2, the rationale for inclusion and decisions on level have focused on the vocabulary that learners around the world seem to know and use. To establish this, we have referred to a range of sources, including written learner data in the Cambridge Learner Corpus, first language corpus data, exam wordlists, and wordlists in coursebooks and other classroom materials. All of the draft entries compiled for the A1–B2 version were reviewed by experienced English language teaching professionals, and other experts were involved in the later validation phase of these levels (see Section 3).

As Section 3 of Capel (2010) suggests, the gap between receptive understanding and productive use at these levels may not be as wide as some people have claimed (see Melka 1997). Modern communicative classrooms encourage far more spoken practice than was the case a generation ago and outside the classroom there are endless opportunities for actively using new language, through mobile technology and the Internet. For this reason, we have not made a distinction between receptive knowledge and productive use up to B2 level.

For the C levels, the methodology is somewhat different (see Section 4 below). Here, receptive knowledge is likely to be broader than actual productive range. Learners will also be using skills of deduction to process unknown words and phrases in context, a strategy that is commonly introduced at the B2 level and is standard practice at the C levels, where there is a need to process large amounts of ungraded text in a field of work or study. Given the domain-specificity at these higher levels and the lack of coursebook wordlists at C1 and C2, we have focused on core vocabulary and have based our research at these levels on actual learner evidence, frequency information from first language corpora, and additional sources for Academic English.

## 3. Validation phase of the A and B levels

The evaluation and validation of the first four levels of the resource aimed to test the usability of the online platform, to verify the decisions taken on CEFR levels and to assess the actual

coverage, with a view to adding anything relevant at A1–B2 that had been inadvertently omitted. To this end, password access was provided to known user groups, notably Cambridge University Press authors, editors and lexicographers, and Cambridge ESOL item writers and exam developers, who worked with the resource over a twelve-month period and submitted detailed comments via the feedback button. These comments were acted on and any apparent level discrepancies were further researched, with revisions often made as a result. An online questionnaire was also completed by these users, which largely focused on the first aim, usability.

Specific validation tasks were carried out by academics based in Tokyo, Miami, Cambridge and Nottingham. In Tokyo, Professor Masashi Negishi and colleagues at Tokyo University of Foreign Studies developed a phrasal verbs test to assess the accuracy of the CEFR levels assigned, which was administered to more than 2,500 students in Japan. This test was also administered to smaller groups of learners in Spain and the Czech Republic.

At Miami Dade University, Dr Michelle Thomas validated the American English version. At the University of Cambridge, Professor John Hawkins and Dr Luna Filipović used the resource during their work on criterial features (Hawkins & Filipović, forthcoming). Cambridge ESOL's Research and Validation expert Dr Angeliki Salamoura carried out quantitative validation research on the A1–B2 data in June 2011, which is described below in Section 4.

Dr Ron Martínez at the University of Nottingham carried out extensive analysis of the phrases in the pilot version, using his own PhD research (Martínez 2011), a list of phrasal expressions based on native-speaker frequency in the British National Corpus. As a result, some 200 'missing' phrases were flagged for possible inclusion, either within the AB levels or at the C levels. Some of these phrases were in fact 'embedded' in dictionary examples for individual senses rather than omitted altogether, but in several cases it was decided to raise their profile by recording them separately. An interesting example of this policy is the phrase *a number of* meaning 'several', which was embedded in the B1 sense AMOUNT and later became a separate phrase entry at B2. A large proportion of the truly missing phrases turned out to be more suited to the C levels and were added to the subsequent compilation process. For further discussion of this aspect of the project, see Section 6 below.

#### 4. Scope of the C levels research

With around 4,700 headwords included up to B2 level, the research team needed to put a provisional figure on the number of additional headwords for C1 and C2. In the context of first language use, Francis and Kucera (1982) analysed the Brown Corpus and found that while a vocabulary size of 5,000 words accounted for 88.7 per cent of the corpus coverage, that figure only rose marginally to 89.9 per cent for 6,000 words (see Schmitt & McCarthy 1997). In the context of second language learning, Adolphs and Schmitt (2003) revisited the research of Schonell et al. (1956) into target vocabulary size for spoken use by analysing the

5-million word CANCODE spoken corpus, and found that knowledge of 5,000 words would be needed to cover 96 per cent of the language in that corpus.

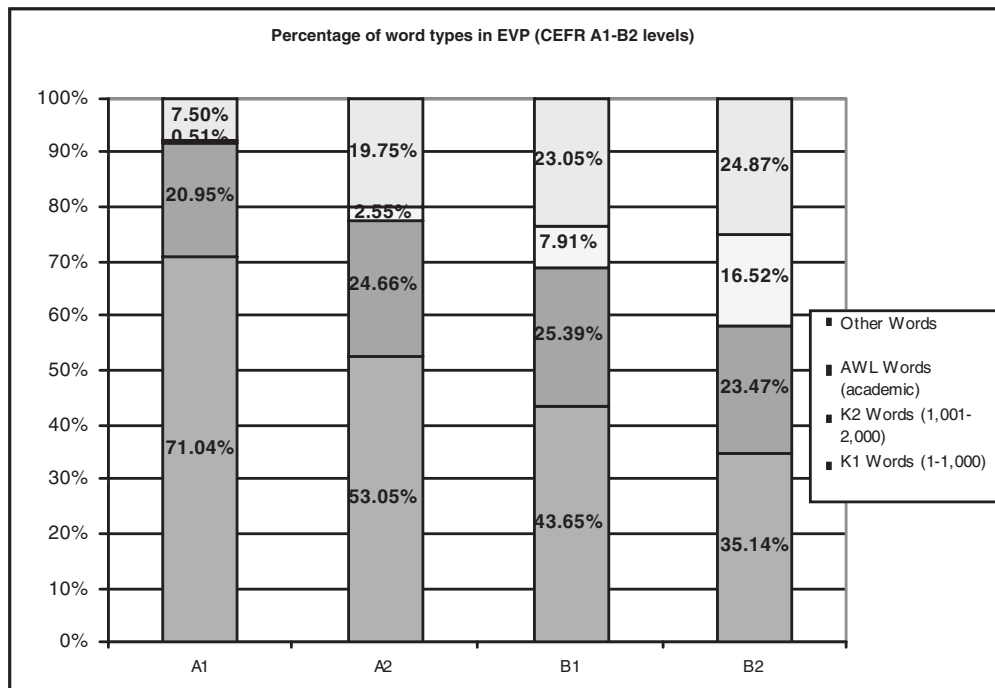
At the outset of the C levels phase of the project, a target of 6,500–7,000 headwords for the complete six-level resource was set, to be refined and determined by actual corpus evidence once the research got under way. As Roland Hindmarsh had done in his *Cambridge English Lexicon* (1980), we also wanted to consider including any remaining senses of the headwords already covered at the AB levels. These less frequent senses of frequent words in English are often crucial to vocabulary development, and in most cases represent meanings and phrases that C-level learners might be expected to know. A preliminary inventory of these additional senses was itemised, with the relevant parts of dictionary entries extracted from the *Cambridge Learner's Dictionary* (Woodford 2007) and the data inserted into the C-level database.

Various sources were used to determine the inclusion of new headwords. As was the case for the A–B levels, dictionary frequency again played its part. Entries for all words tagged I and A in the *Cambridge Advanced Learner's Dictionary* that had not been included up to B2 level were added to the C-level database, along with additional words derived from learner corpus evidence. In addition, words were taken from the Academic Word List (Coxhead 2000), to be checked against learner evidence before inclusion. Almost all of the most frequent family members listed in italics in the ten sub-lists of the Academic Word List have been included, with one or two exceptions that were considered too specialised and for which there was no learner evidence, for example the word *protocol*.

For the C levels, it was decided that both Academic English and General English should be covered, and consequently the learner corpora consulted included the International English Language Testing System (IELTS) data. This proved to be quite a challenge, because unlike the general English ESOL exams, IELTS reports at different levels of ability. An ‘academic’ learner seeking entry into a university might have followed an IELTS preparation course pitched at the C1-level threshold (IELTS 6.5–7) and have ‘learned’ C1-level vocabulary as part of this preparation, but fail to achieve higher than B2 level in the IELTS exam itself. Therefore, new headwords seen as likely for inclusion at C1 might appear as frequent at B2 level or below in the IELTS data, but this could well be due to underachievement rather than positive ‘can do’ ability. A pragmatic and common-sense approach was taken here, verifying CEFR level through other sources where possible.

There is now a considerable amount of C1- and C2-level data in the Cambridge Learner Corpus. We reviewed frequency-ordered lists of the words in the Cambridge English: Advanced (CAE), Cambridge English: Proficiency (CPE) and IELTS data and came to the conclusion that in order to confidently include a new headword in the English Vocabulary Profile (EVP) there should be multiple instances of use, across more than one exam session. Our minimum number of raw occurrences for any potential new C-level headword was set at fourteen, using in the first instance the CAE learner data for C1 and the CPE data for C2, with raw frequencies in the IELTS data also checked. Words used on the question paper often provided falsely inflated figures, especially in the IELTS data – for example, the word *deforestation* was one of the highest frequencies listed, but only due to its use on the question paper in a Part 1 task on one exam session.

At the time of writing, there are 6,970 headwords in the EVP resource from A1 to C2 levels. However, although the addition of new words is fairly modest for the reasons explained



**Figure 1** "Percentage of word types in EVP (CEFR A1-B2 levels)".

above, there are around 15,000 senses and phrases listed for the A1–C2 resource, and at the C levels the increase in the number of senses and phrases amounts to more than 5,000. For more details on this aspect, see Section 5 on polysemous words below.

The Academic Word List was an important new source for the C-levels research, providing as it does a listing of the most frequent words used in academic text. These are grouped as word families in ten sub-sets by frequency. We took the most frequent form of each family in all ten sub-sets and looked for evidence of it in the C-level learner data, with a view to including all such words in the six-level resource. Of course, some were already featuring at the lower levels. As Dr Angeliki Salamoura’s (2011) research has shown, three words even appear at A1 level – *adult*, *computer*, *job* – but arguably these are words that operate in general English as well as academic English. Figure 1 illustrates Salamoura’s findings in June 2011, where she compared the A1–B2 data with Lextutor’s top 2,000 words for first language use (split into K1 and K2; see Cobb, no date) and the Academic Word List. She is now conducting similar research on the C1- and C2-level data.

Interestingly, the fourth category, ‘other words’, remains a sizeable component across the A2–B2 levels and represents all those words that are less frequent but which are important to learners, because they belong to topics which they are interested in – *download* at A2, for example.

Returning to academic words and phrases, another source that proved very informative was the Academic Formulas List, researched by Nick Ellis and Rita Simpson-Vlach (2010), to which we were given early pre-publication access. Effectively a phrase list for Academic English, this suggested important collocates, which we were able to highlight in the C-level dictionary examples, and also certain academic phrases, which were cross-checked with our other sources, most notably the work of Martínez.<sup>1</sup>

Concluding this section on the ‘scope’ of the C-levels research, it has to be acknowledged that users accessing these final two levels of the English Vocabulary Profile will inevitably find ‘omissions’. Learners who reach this advanced stage will be acquiring vocabulary that is relevant to their specific domain of study, work or interest, and it is beyond the remit of the resource to be exhaustive in that way. The C levels of the resource are a reflection of typical learner vocabulary in the areas of General and Academic English, and are not intended to offer a ready-made lexical syllabus. What has been sought is a common core, in order to describe the words, phrases, phrasal verbs and idioms that learners know and, in most cases, can use (see Capel [2010] for a discussion of ‘knowing’ and ‘using’) – we have included a few frequent senses, phrases and idioms without supporting evidence from the Cambridge Learner Corpus, in cases where our reviewers have supported their inclusion and have confirmed that such lexis is likely to be known by C2 level. As stated at the outset of this article, our learner data needs adding to, especially spoken learner language.

## 5. Polysemous words

At the outset of the project in 2007, a database of dictionary entries was made available to the research team, taken from the *Cambridge Advanced Learner’s Dictionary*. This data supplied reliable frequency information for first language use for the individual meanings of a word, based on the actual counting of corpus lines by meaning. Although all good monolingual dictionaries include frequency information at headword level, no other monolingual dictionary has frequency information for individual meanings, and it was immediately apparent how useful this would be to the project as a starting point.

In searching for learner evidence of the different meanings of these frequent words, some interesting findings emerged. As already discussed in Capel (2010) on the A1–B2 levels, the most frequent sense for first language users is not always the first to be taught, and there are often sound reasons for this. The acquisition of concrete meanings tends to precede more abstract ones, so the meanings of the words *case* and *stage* that are taught at A2 level are the physical ones (‘pencil cases’ and ‘the raised area for acting on’), while the most frequent meanings of these two nouns only appear to be known at the B levels – the meaning of *case* SITUATION is B1 and the meaning of *stage* PART, ‘a period of development’, is B2. Furthermore, our review of wordlists in course materials indicated that the most frequent meanings are sometimes never taught explicitly, as in *case* SITUATION.

<sup>1</sup> See <http://nottingham.academia.edu/Rmartinez>.

The English Vocabulary Profile uses capitalized guidewords as just illustrated in order to make it easier to navigate the very long entries for words with multiple meanings: there are 33 matches for the noun *way*, 83 for the verb *go* and 109 for the preposition *at*. Long entries will usually include a number of phrases (see Section 6), and may also include phrasal verbs and idioms, making it all the more important to highlight the distinct meanings of a word clearly. For the end users of the resource, whether they are materials writers, exam setters, teachers or students, the guidewords provide a swift summary of the scope of learner knowledge at each CEFR level, which can inform teaching/learning priorities.

When it came to researching the C levels we found that, occasionally, some of the less frequent senses of words that had already been included in the A1–B2 levels of the EVP failed to make any appearance in the written learner evidence of the Cambridge Learner Corpus at the C levels. As these words are generally within the 5,000 most frequent for first language use, and the senses that were under consideration are included in the *Cambridge Learner's Dictionary* (aimed at intermediate learners of English), we were reluctant to instigate a blanket exclusion policy purely on the basis of lack of evidence in the Cambridge Learner Corpus (CLC).<sup>2</sup> There were indeed arguments for the inclusion of remaining senses if only for the sake of completeness in terms of the EVP resource – words that had already made it into the A1–B2 levels were clearly important for learners and a full picture of their multiple meanings would probably be of benefit in the language classroom.

Accordingly, we investigated senses without CLC evidence further. In most cases, the lack of learner examples could be explained either by the nature of the exam tasks set or because of their predominantly spoken use. A recommendation was usually made to include these instances at C2 unless outside expert opinion argued against this. A small sub-set of ‘colloquial’ spoken senses, as in the sense GOOD of *mean* in Table 2, were omitted from the resource but still remain as ‘suppressed’ senses in our internal database, which can be revisited once spoken learner data is available.

The deciding factors for any other ‘missing’ senses were their relative frequency in first language use, how specialised they are in meaning, and how far they belong to General English as opposed to other domains, such as Business English. Table 2 illustrates this thought process and gives some examples of decisions made.

A representative selection of additional meanings that have been included at the C levels and ones that have been omitted are given in Tables 1 and 2. Note that this does not cover meanings/uses presented as phrases (see Section 6).

## 6. Phrases in the English Vocabulary Profile

Many of the remaining senses from the *Cambridge Learners Dictionary* data for consideration at the C levels were in fact presented as phrases rather than meanings with guidewords. This ties in the greater focus on semi-fixed phrases and collocations in the advanced language classroom and, for the most part, learner evidence was found to justify the inclusion of these

<sup>2</sup> See [http://www.cambridge.org/gb/elt/catalogue/subject/custom/item3646603/Cambridge-English-Corpus-Cambridge-Learner-Corpus/?site\\_locale=en\\_GB](http://www.cambridge.org/gb/elt/catalogue/subject/custom/item3646603/Cambridge-English-Corpus-Cambridge-Learner-Corpus/?site_locale=en_GB).

Table 1 *Additional meanings of polysemous words included at the C levels*

Headword	Meanings at A1–B2 levels	Meanings at C levels	Learner example for C level entry?
click (verb)	COMPUTER (A2)	IDEA (C2)	Yes
		SOUND (C2)	Yes
		PEOPLE (C2)	Yes
force (noun)	POWER (B2) GROUP (B2)	INFLUENCE (C2)	Yes
		Memory	COMPUTING (A2) ABILITY TO REMEMBER (B1) EVENT REMEMBERED (B1)
plain (adjective)	SIMPLE (B1) NOT MIXED (B1)	OBVIOUS (C2)	Yes
		PERSON (C2)	Yes
rough	NOT SMOOTH (B1)	DANGEROUS (C1)	Yes
	NOT EXACT (B1)		
	SEA/WEATHER (B2)		
	DIFFICULT (B2)		
Serve	PROVIDE FOOD/DRINK (A2)	BE USEFUL (C1)	Yes
	SHOP (B1)	WORK (C1)	Yes
	TIME (B1)	PRISON (C2)	Yes
Within	DISTANCE (B1)	INSIDE (C1)	Yes
	LIMIT (B2)		

phrases. As explained above, however, inclusion did not rest on the CLC alone. So, for example, the complete entry for the adjective *sharp* includes three phrases at C2 level:

*a sharp pain*  
*a sharp bend/turn, etc.*  
*a sharp contrast/difference, etc.* – a very big and noticeable difference between two things

The last of the three phrases above has no learner example, but it was seen as a key phrase to include, especially for Academic English, and is likely to be known at C2 level – the core meaning of *contrast* in this phrase, DIFFERENCE, is already included at B2 in the EVP. Pragmatic decisions such as this were not reached internally, but in consultation with outside expert reviewers.

Thanks to the validation work carried out by Martínez on the A1–B2 levels, we had a further list of possible phrases to include at the C levels, drawn from his research into frequent phrasal expressions. All of these phrases have been included in the EVP. As Martínez has so clearly demonstrated in talks and papers, (and see Martínez, no date) even if learners know the top 2,000 words in English, the use of these words in phrases will not always be grasped, particularly when the meaning of the phrase as a whole is more figurative. The EVP research



Table 2 *Less frequent meanings of polysemous words omitted from EVP*

Headword	Meanings included in EVP	Meaning(s) omitted	Decision
Capacity	AMOUNT (B2) ABILITY (B2) JOB (C1)	PRODUCE (the amount that a factory or machine can produce)	More relevant to Business English
file (noun)	COMPUTER (A2) INFORMATION (A2)	TOOL	Specialist use
Indicate	SAY (B2) SHOW (C1)	SIGNAL (when driving)	UK and specialist use
mean (adjective)	NOT GENEROUS (B2) NOT KIND (B2) AVERAGE (C2)	GOOD (informal use, always before noun – <i>I make a mean spaghetti.</i> )	Colloquial sense – revisit once there is spoken learner data available
plug (noun)	ELECTRICAL OBJECT (B1) HOLE (B1)	ADVERTISEMENT	More relevant to Business English
Racket	SPORT (A2) NOISE (C2)	ILLEGAL ACTIVITY	Infrequent sense
skin (noun)	BODY (B1) FRUIT (B2)	LIQUID (a thin, solid layer on top of a liquid) COMPUTER (the way information is arranged on a computer screen)	Infrequent sense Specialist use

has borne this out and many phrases formed from very frequent words are nevertheless listed at the C levels. For example, although the word *short* first appears as an adjective at A1, its other parts of speech are confined to the C levels: the noun appears in the phrase *in short* at C1, and the adverb occurs in several phrases, including *to cut a long story short* at C1 and *fall short of something*, which has two distinct meanings at C2.

At the B levels, learners do not appear to be using even relatively transparent phrases formed from very frequent words as early as might be expected – for example, the phrase *a number of*, meaning ‘several’, is given B2 in the resource on the basis of current source evidence. Of course, this phrase would probably be readily understood from its surrounding context at a lower level, but from our research it seems that learners are not meeting it overtly in the classroom until a later stage. Given the usefulness and frequency of this phrase, it could possibly be taught earlier, and this is true of many other frequent phrases. However, throughout the A and B levels, there still seems to be a heavy emphasis on the teaching of topic vocabulary, largely because coursebook units tend to be organised according to topic coverage. This often leads to repetition across coursebooks at the B levels, and seems an inefficient way of offering vocabulary development. In contrast, from B1 onwards, the teaching syllabus might usefully introduce areas of vocabulary that are less tied to topics, including frequent phrases like *a number of*.

Through its advanced search menu, the EVP can provide instant checklists of useful phrases at each CEFR level – the number of phrases currently included at each level is given in Table 3. The numbers of phrases known at A1 and A2 are understandably modest, but at

Table 3 *Phrases listed at each level of the EVP*

A1	47
A2	147
B1	335
B2	561
C1	380
C2	722

Table 4 *Phrasal verbs listed at each level of the EVP*

A1	4
A2	27
B1	135
B2	277
C1	99
C2	184

B1 and B2 there is a noticeable increase. What is slightly surprising is the apparent drop at the C1 level, though the total number of phrases across the two C levels (1,102) is a higher figure than B1 and B2 combined (896). It is also worth saying that the advanced search for ‘words’ that are new at C1 level does give a sizeable figure: the total for new words and additional meanings at C1 is 1,622. We will continue to seek feedback on the resource in order to validate these findings.

It is interesting to see what phrases are typical of the B levels and what kind of phrases learners add at the C levels. If a phrases search is conducted at two different CEFR levels, the results are interesting. For example, there are many prepositional phrases at B1 – *at once, by mistake, for fun, in fact, on purpose, out of order* – as well as phrases for informal communication, such as *a load / loads, be into something, be up to something, no way, too bad, what’s up?* In contrast, there are several new phrases at C1 that are formal – *in accordance with something, feel compelled to do something, be glad of something, sincere apologies, as yet* – and many phrases that are relevant to academic writing, such as *it would appear (that), when it comes to something / doing something, to the contrary, in any event, first and foremost, on the grounds of / that, be inclined to think / believe / agree, etc., by and large, for the most part, regardless of, safe to say*. Making comparisons like this might help users of the EVP resource to orientate themselves for a particular group of learners: teachers about to start a new class at an unfamiliar level, for example.

Note that phrasal verbs are presented as a separate category in the EVP, so it is possible to search for these in isolation. As reported in Capel (2010), researching the level of phrasal verbs has been problematic, due to the relatively small amount of CLC evidence – learners often lack confidence in this area and may avoid using phrasal verbs, especially under exam conditions. Until we have access to spoken learner data and the new non-exam written learner corpus, this will remain an issue, and even then, high numbers of learner examples cannot be guaranteed.

The current figures for the number of phrasal verbs included at each level of EVP are given in Table 4. Individual meanings of the same phrasal verb have been counted separately.

## 7. Idioms in the English Vocabulary Profile

Idioms were understandably seen as a research priority for the C levels part of the project. As already stated above, figurative meanings and uses characterise the C2 level in particular and idioms are an extension of this vocabulary area. The research carried out by Martínez (2011) suggested examples of combinations of highly frequent words that are not readily understood due to their idiomatic nature, such as *miss the boat*.

In developing a strategy for the inclusion of idioms in the EVP, we decided on two basic criteria: frequency of current use and CLC evidence. Our important reference sources for first language use were the *Collins COBUILD Dictionary of English Idioms* (Collins 2011) and the *Cambridge Idioms Dictionary* (Walter 2006). Both titles are based on corpus data and the COBUILD one indicates frequency through a three-tier arrow system, with three arrows being the most frequent. The Cambridge dictionary highlights in blue boxes idioms that are “very common and useful to learn”. Halfway through the research period, two new publications from Oxford University Press were added to our checking sources: *Idioms and Phrasal Verbs Intermediate* and *Advanced* (Gairns & Redman 2011a, 2011b). Although these titles do not appear to be corpus-informed, they were very useful in respect of perceived learner level.

The Cambridge Learner Corpus yielded some interesting evidence, though we had to be careful to check the age of the learner examples: a huge number of examples appeared for the old-fashioned idiom *raining cats and dogs*, but on closer examination, all of them were produced in 1993! In any case, this idiom did not match our inclusion criterion of being current in first language use, and so has not been included. On the other hand, the relatively recent idiom *behind closed doors* has been included, at C2 level. This idiom seems to have a number of direct translations into other languages, so it may be that the assigned level is unduly high – however, this is where we have found the learner evidence.

In terms of overall figures, there are 12 idioms at B2, 39 at C1 and 196 at C2. This then is a fairly small total overall, with only 247 idioms currently listed in the EVP. As with phrasal verbs, learners may not be confident in using idioms even at C2 level, and this could account for our modest inclusion. It is an interesting area to consider from the classroom perspective and raises certain questions: Is it important for C2-level learners to be familiar with a wide range of idioms and how many of these idioms should they master in terms of productive ability? This project welcomes your opinions on these questions and any other aspect of the resource – please submit your views using the EVP Feedback facility, situated in the blue panel.

## 8. What is C2 Mastery?

In preparing the C1 and C2 levels of the resource, a huge amount of learner evidence was scrutinised, and there was some extremely impressive writing, given it was done under exam

conditions, particularly at the C2 level. The CEFR term for this level is ‘Mastery’, but what does this actually mean? It is not necessarily the very top performance – we know that some exceptionally gifted learners operate well above the C2 level. Indeed, quite often in our consultation process with outside reviewers, words and phrases without learner evidence were deemed to be above C2, referred to unofficially as ‘D1’! And, at one time, there was in fact a Cambridge exam for this post-Proficiency level, called the Diploma in English.

So, what are the characteristics of Mastery in English? Clearly, accuracy is a factor, with virtually flawless production of complex structures and precision in vocabulary choice. Another important aspect to be considered is that of appropriacy – use of style, register and tone according to genre and audience – but perhaps the best indicator of Mastery is the sophisticated use of a wide lexical range, embracing the different categories of vocabulary that we have been looking at here: words, phrases, phrasal verbs and idioms. Let us consider some C2 learner examples illustrating these categories, all of which feature in the EVP. The words in bold show which entry they appear at:

*We set off, **armed with** all our cameras, lenses, travelling gadgets and equipment.*  
*Suddenly the moon disappeared behind the clouds and, in a few moments, a violent storm **broke**.*  
*In the world we live in today, jobs have become much more difficult to **come by**.*  
*All of a sudden she **caught my eye** and smiled in a sad way.*  
*This very fact made my father work as a slave, as he was the only breadwinner at home, my mother **having her hands full** with us four.*  
*To return to our first point, people do not usually **land a job** in their field of study or childhood dream.*  
*She was always yearning for things **beyond her reach**.*  
*I couldn't **reconcile myself to** the thought that my sister had proved to be smarter than me once again.*  
*I could not understand how these words had **slipped out** of my mouth.*  
*Opposite and above us towered huge mountains like rocky giants reaching their hands up into the cloudless sky. Our painful legs were forgotten, the scratches paled into insignificance in the face of such majestic **splendour**.*  
*We are loyal readers of your newspaper and we ask you to raise your voice in defence of our community against the **unscrupulous** sharks of big business.*  
*It remains to us to prove that the opinions of some scientists are far-fetched and **don't hold water**.*

You will find many other learner examples of a similar standard in the EVP, along with C2 examples that do contain the occasional error (corrected in our examples within square brackets).

## 9. Affixation and the word family panels

Finally, affixation should be mentioned in the context of the C-levels work. As explained in Capel (2010), for the A1–B2 levels it was decided to separate out dictionary ‘run-ons’ and make them entries in their own right, provided these words were sufficiently frequent in first language use and were supported by learner evidence. In the case of certain adverbs, which seem to be somewhat underused by learners, we occasionally allowed entries to stand without a learner example.

Word family panels were then created in order to display related words together, and these appear at the top of each headword entry that belongs to a family of at least two words. Clicking on any word within a panel takes the user to the entry for that word.

When it came to the C levels, we wanted to distinguish existing word family members from the more advanced additions at C1 and C2, and decided to do this typographically, putting the C level words into italics (as shown below). In this way, teachers and materials writers can see at a glance how learners typically extend their knowledge of word families, and test setters can use the panels to target specific forms at an appropriate CEFR level.

### **different**

#### **Word family:**

**Nouns:** difference, *indifference*

**Verbs:** differ, *differentiate*

**Adjectives:** different, *indifferent*

**Adverbs:** differently

We welcome feedback on the word panels and would love to hear how they are actually being used.

### **Acknowledgements**

The development of the English Vocabulary Profile in its present form is the result of an extremely talented, creative and dedicated team of people, who have expertise in many different areas. Special thanks go to Liz Walter for her outstanding lexicographic support, to Daniel Perrett for his extraordinary computing skills, to Dorota Bednarczyk-Krajewska for her highly competent management of the electronic development of the resource, to Carol-June Cassidy for her painstaking collaboration on the US version, and to Helen Naylor for her expert judgement on CEFR levels, derived from a lifetime of classroom experience.

### **References**

- Adolphs, S. & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics* 24.4, 425–438.
- Capel, A. (2010). A1–B2 vocabulary: Insights and issues arising from the English Profile Wordlists. *English Profile Journal* 1.1.
- Cobb, T. (no date). Web Vocabprofile. <http://www.lex Tutor.ca/vp/> (accessed 6/2011).
- Collins COBUILD dictionary of English idioms* (2011). 2nd edn.
- Coxhead, A. (2000). The Academic Word List. <http://www.victoria.ac.nz/lals/resources/academicwordlist/default.aspx>.
- Ellis, N. & Simpson-Vlach, R. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31.4, 487–512.
- Francis, W. N. & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.
- Gairns, R. & Redman, S. (2011a). *Oxford word skills: Idioms and phrasal verbs intermediate*. Oxford: Oxford University Press.
- Gairns, R. & Redman, S. (2011b). *Oxford word skills: Idioms and phrasal verbs advanced*. Oxford: Oxford University Press.
- Hawkins, J. A. & Filipović, L. (forthcoming). *Criterial features in L2 English*. Cambridge: Cambridge University Press.
- Hindmarsh, R. (1980). *Cambridge English lexicon*. Cambridge: Cambridge University Press.

- Martínez, R. (2011). *The development of a corpus-informed list of formulaic sequences for language pedagogy*. Unpublished PhD thesis, University of Nottingham.
- Martínez, R. (no date). <http://nottingham.academia.edu/Rmartinez>.
- Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In Schmitt, N. & McCarthy, M. (eds.), *Vocabulary: Description, acquisition, and pedagogy*. Cambridge: Cambridge University Press, 84–102.
- Salamoura, A. (2011). Unpublished paper presented at the ALTE Conference.
- Schmitt, N. & McCarthy, M. (eds.) (1997). *Vocabulary: Description, acquisition, and pedagogy*. Cambridge: Cambridge University Press.
- Schonell, F. J., Meddleton, I. G. and Shaw, B. A. (1956). *A study of the oral vocabulary of adults*. Brisbane: University of Queensland Press.
- Walter, E. (ed) (2006). *Cambridge idioms dictionary*, 2nd edn. Cambridge: Cambridge University Press.
- Walter, E. (ed) (2007). *Cambridge learner's dictionary*, 3rd edn. Cambridge: Cambridge University Press.